Lecture 15. *Information Theory*

- We will learn that information should be treated as a physical property

https://learn.g2.com/what-is-information-technology

What is information?

Consider this: Thanksgiving is a US holiday and we want to know when it takes place.

- **Let's compare three different statements**
  - Thanksgiving falls on a particular day of the year → No information!
  - Thanksgiving falls in the last quarter of the year → More information!
  - Thanksgiving falls on a Thursday → Even more information!
- **How much information do we obtain from each of those statements?**

We have an intuitive understanding of the amount of information but How can we quantify this?

Quantifying information

Let's look at probabilities

- Thanksgiving falls on a particular day of the year → P=1
- Thanksgiving falls in the last quarter of the year → P=1/4
- Thanksgiving falls on a Thursday → P=52/365

If you were to gamble by guessing, you would be better off knowing the last statement!

We conclude that there is more information in that statement

It appears that the information is related to the *inverse* of the probability

What if you know two statements?

- Thanksgiving falls on a particular day of the year → P=1
- Thanksgiving falls in the last quarter of the year → P=1/4
- Thanksgiving falls on a Thursday → P=52/365

Since the statements are independent, knowing two statements ***increases*** the chances by **multiplying** the probabilities.

In that case the information is ***added***

It looks like a logarithm should show in the definition of information!

Claude Shannon definition of information

$$Q = -k \log P$$

Q is the amount of information, measured in *bits*

k is a positive constant

Claude Shannon entropy

Average information

$$S = \langle Q \rangle = \sum_i Q_i P_i = -k \sum_i P_i \log P_i$$

For a set of statements with probability P_i

$$S = -k \sum_i P_i \log P_i$$

According to Shannon's definition:
The lower the probability (the higher the uncertainty), the more information we gain!

Example: you play roulette, is it better to know the color or the number?

Shanon entropy of a Bernoulli trial

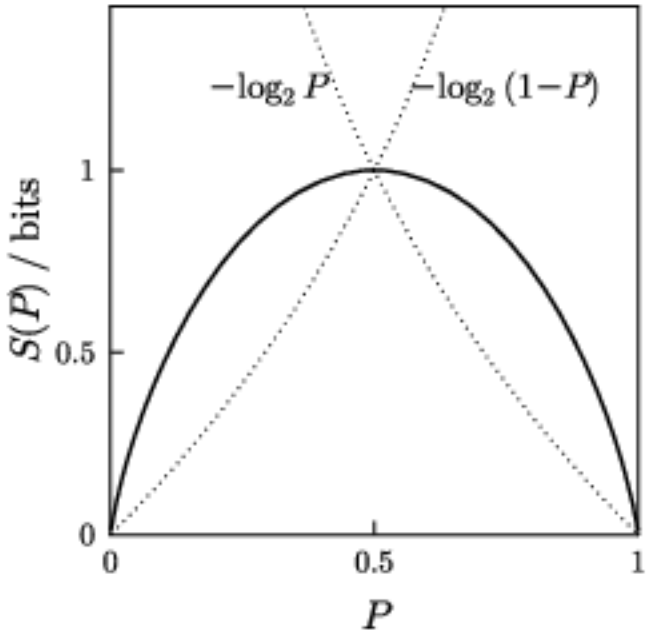
Bernoulli trial: a 2-outcome experiment where one outcome occurs with probability P

$$S = -\sum_i P_i \log P_i = -P \log P - (1-P)\log(1-P)$$

- Entropy is maximal for P=1/2 (most uncertainty)
- Entropy is minimal at P=0 or P=1 (no information)

Example: compare the information you can obtain from the following two Bernoulli trials.

- Physics students like math with probability P and don't like math with probability 1-P
- Physics students like sautéed shrimps with probability P and don't like them with probability 1-P



Information and thermodynamics

$$S = \langle Q \rangle = \sum_i Q_i P_i = -k \sum_i P_i \log P_i$$

It looks identical to Gibbs' expression for thermodynamic entropy!

It is a measure of uncertainty, based on its properties (macrostates) but limited knowledge of its microstates.

Entropy and information are closely connected!

Rolf Landauer even claimed that information ***is*** a physical quantity!

Information is a physical quantity

Rolf Landauer claimed that information is a physical quantity!

Imagine you have N bits of information, connected to a thermal reservoir at temperature T

Now we erase that information. This process is **irreversible**.

This is done by changing all the bits to zero

This process reduces the number of (micro-) states by 2^N

The entropy goes **down** by $Nk_B \ln 2$ or $k_B \ln 2$ per bit

The entropy of the universe cannot decrease so it must go up by $k_B \ln 2$ by bit

We must dissipate heat in the surroundings equal to $k_B T \ln 2$ per bit erased

This provides a resolution to Maxwell's demon problem we discussed earlier!

Data compression

If you have *N* bits, you would think you have the same information regardless of their distribution.

What is the *real* information stored in a particular block of data?

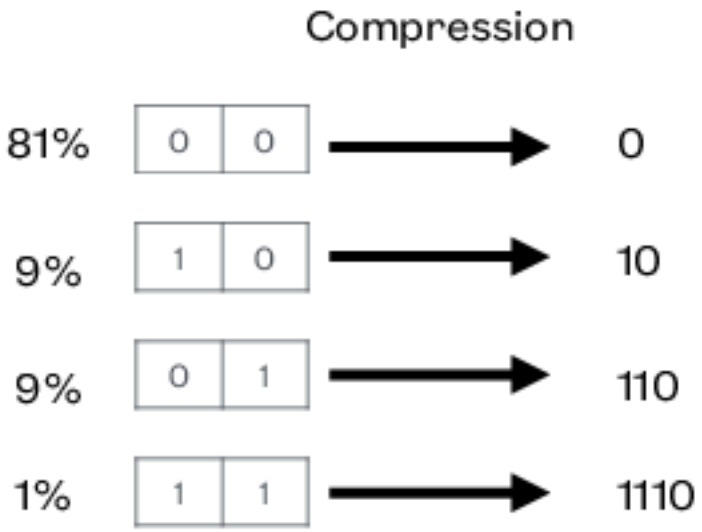
Example, the English language includes many occurrences of the termination **-ing**

Could this be compressed as a one bit of information? This is the idea behind data compression

This data compression idea is formalized in **Shannon's noiseless channel coding theorem.**

We will not prove it or study it in depth but we will introduce it here

Data compression



We save one bit in the most likely case!

Imagine that the first bit is 0 with probability 0.9 and the second bit is 0 with probability 0.9 as well

Imagine you have 100 bits, you know, on average that 81 of them are 00, 9 are 10, 9 are 01 and 1 is 11, the total number of bits used is thus:

$$81 \times 1 + 9 \times 2 + 9 \times 3 + 1 \times 4 = 130$$

Using the uncompressed scheme: $100 \times 2 = 200$

More information requires more bits!

We must identify typical sequences and efficiently code only those!

Imagine you have some data, and you divide them into sequences of length *n*.

We also imagine that the data points are uncorrelated (that is: independent)

Probability of finding a sequence $x_1, x_2, \ldots, x_n$

$$P(x_1, x_2, \ldots, x_n) = P(x_1)P(x_2)\ldots P(x_n) \approx P^{nP}(1-P)^{n(1-P)}$$ for a typical sequence.

$$-\log_2 P(x_1, x_2, \ldots, x_n) \approx -nP\log_2 P - n(1-P)\log_2(1-P) = nS$$

$$\longrightarrow \quad P(x_1, x_2, \ldots, x_n) \approx \frac{1}{2^{nS}}$$

$$\longrightarrow \quad P(x_1, x_2, \ldots, x_n) \approx \frac{1}{2^{nS}}$$

We have at most 2^{nS} typical sequences and we require nS bits to code them

For large enough n, the typical sequences become longer and the probability of this compression scheme to fail becomes smaller.

Conditional and joint probabilities

Probability of an event usually depends on what happened before

More information will modify the probability of knowing other information.

Conditional probability $P(A|B)$

Probability that event A occurs given than event B has happened

Joint probability $P(A \cap B)$

Probability that events A and B occur

We see directly that: $P(\mathrm{A} \cap \mathrm{B}) = P(\mathrm{A}|\mathrm{B})P(\mathrm{B})$ and $P(\mathrm{A} \cap \mathrm{B}) = P(\mathrm{B}|\mathrm{A})P(\mathrm{A})$

Independent events

$$P(\mathrm{A} \cap \mathrm{B}) = P(\mathrm{A}|\mathrm{B})P(\mathrm{B})$$

$$P(\mathrm{A}|\mathrm{B}) = P(\mathrm{A})$$

(it does not matter if we know something about B or not!)

$$P(\mathrm{A} \cap \mathrm{B}) = P(\mathrm{A})P(\mathrm{B})$$

Set of mutually exclusive events A_i

$$\sum_i P(\mathrm{A}_i) = 1 \quad \longrightarrow \quad P(\mathrm{X}) = \sum_i P(\mathrm{X}|\mathrm{A}_i) P(\mathrm{A}_i)$$

$$P(\mathrm{A} \cap \mathrm{B}) = P(\mathrm{A}|\mathrm{B})P(\mathrm{B})$$

$$P(\mathrm{A} \cap \mathrm{B}) = P(\mathrm{B}|\mathrm{A})P(\mathrm{A})$$

We also know that: $P(\mathrm{A} \cap \mathrm{B}) = P(\mathrm{B} \cap \mathrm{A})$

Thus: $P(\mathrm{B} \mid \mathrm{A})P(\mathrm{A}) = P(\mathrm{A} \mid \mathrm{B})P(\mathrm{B})$

$$P(\mathrm{A}|\mathrm{B}) = \frac{P(\mathrm{B}|\mathrm{A})P(\mathrm{A})}{P(\mathrm{B})}$$

Bayes' theorem

P(A) is the prior probability

Bayes theorem yields the posterior probability

Bayes' theorem: example

Imagine you have a large campus and you perform a covid-19 test.

Probability that a student has covid-19 is 5% and the probability a student does not is 95%

$$P(C) = 0.05 \qquad P(\bar{C}) = 0.95$$

Imagine that we test for covid-19. Suppose the test is not fully reliable but only p% of all test are accurate.

What is the probability that a student has a positive test?

$P(T\|C)$	Probability to be tested + if infected
$P(T\|\bar{C})$	Probability to be tested + if not infected
$P(\bar{T}\|C)$	Probability to not be tested + if infected
$P(\bar{T}\|\bar{C})$	Probability to not be tested + if not infected

$P(T|C) = p$ Probability to be tested positive if infected

$P(T|\bar{C}) = 1 - p$ Probability to be tested positive if not infected

$P(\bar{T}|C) = 1 - p$ Probability to not be tested positive if infected

$P(\bar{T}|\bar{C}) = p$ Probability to not be tested positive if not infected

Probability to have a positive test?

$$P(T) = P(T|C)P(C) + P(T|\bar{C})P(\bar{C})$$

$$P(T) = p \times 0.05 + (1-p) \times 0.95 = 0.95 - 0.9p$$

If p =0.99, then we have P(T)=0.059

Probability that a student has covid-19 if tested positive?

$$P(C|T) = \frac{P(T|C)P(C)}{P(T)} \qquad P(C|T) = \frac{p * 0.05}{0.95 - 0.9p}$$

p=0.99 -> P(C|T)=84%

Probability that a student has covid-19 if tested negative?

$$P(C|\bar{T}) = \frac{(1-p) * 0.05}{0.05 + 0.9p}$$

$$P(C|\bar{T}) = 0.05\,\%$$

A more intriguing example

Mrs Trellis (from North Wales) has two children, born three years apart. One of them is a boy.

What is the probability Mrs Trellis has a daughter?

Mrs Trellis has two children and the taller of her children was a boy

What is the probability Mrs Trellis has a daughter?

Probability depends on the information we know!

1) One of the children is a boy
2) No other useful information!

Three possible scenarios:
(ranked in order of age)

(1) Boy, boy

(2) Boy, girl

(3) Girl, boy

What is the probability of each outcome?

1/3!

Thus, the probability that Mrs. Treillis has a daughter is 2/3!!!!!!

Mrs Trellis (from North Wales) has two children. The youngest of them is a boy.

What is the probability Mrs Trellis has a daughter?

Thus, probability that Mrs. Treillis has a daughter is 1/2!!!!!!

Mrs Trellis has two children and the taller of her children was a boy

What is the probability Mrs Trellis has a daughter?

Two possible scenarios: (order of height)

(1) Boy, boy

(2) Boy, girl

Thus, the probability that Mrs. Treillis has a daughter is 1/2

We will see the impact of this when we treat the indistinguishability of particles and their statistics.

Information theory provides a rationale for setting up probability distributions on the basis of partial knowledge; one simply maximizes the entropy of the distribution subject to the constraints provided by the data.

Thermodynamics also gives the best description of the properties of a system that has so many $\left(\approx 10^{23} \right)$particles that one cannot follow it precisely; the Boltzmann probability obtained by maximizing the Gibbs entropy is the least-biased estimate of the probability consistent with the constraint that a system has fixed internal energy U.

Summary

The information Q is given by $Q = -\ln P$ where P is the probability.

The entropy is the average information $S = \langle Q \rangle = -\sum_i P_i \log P_i$

Bayes' theorem relates the posterior probability (which is a conditional probability) to the prior probability.